AD_____

MIPR NO: 93MM3513

TITLE: PROPOSAL FOR RESEARCH IN QUANTITATIVE BIOASSAY METHODOLOGY AND RISK ANALYSIS AND CHARACTERIZATION

PRINCIPAL INVESTIGATOR: Donald P. Gaver

CONTRACTING ORGANIZATION: Naval Postgraduate School
Code 21
Monterey, California 93943-5000

REPORT DATE: September 29, 1994

TYPE OF REPORT: Annual Report

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick
Frederick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for public release;
distribution unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 29, 1994	Annual (1/11/93 - 1/11/94)

4. TITLE AND SUBTITLE
Proposal for Research in Quantitative Bioassay Methodology and Risk Analysis and Characterization

5. FUNDING NUMBERS
MIPR No.
93MM3513

6. AUTHOR(S)
Donald P. Gaver
Patricia A. Jacobs

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School, Code 21
Monterey, California 93943-5000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick
Frederick, Maryland 21702-5012

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release;
distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

A summary of research conducted during the period January 11, 1993 - January 11, 1994, is given. Included are the results of a simulation study of the behavior of estimators of the teratogenic index; a statistical procedure to assess between tank variability and the results of its use on data from a bioassay experiment with medaka; and procedures to combine the results of different experiments using different biomarkers.

14. SUBJECT TERMS
Teratogenic Index, combining information, assessment of between-tank variability

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to *stay within the lines* to meet *optical scanning requirements*.

Block 1. Agency Use Only *(Leave blank)*.

Block 2. Report Date. Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least the year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C	- Contract	PR	- Project
G	- Grant	TA	- Task
PE	- Program Element	WU	- Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es). Self-explanatory.

Block 10. Sponsoring/Monitoring Agency Report Number. *(If known)*

Block 11. Supplementary Notes. Enter information not included elsewhere such as: Prepared in cooperation with...; Trans. of...; To be published in.... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR).

DOD - See DoDD 5230.24, "Distribution Statements on Technical Documents."
DOE - See authorities.
NASA - See Handbook NHB 2200.2.
NTIS - Leave blank.

Block 12b. Distribution Code.

DOD - Leave blank.
DOE - Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.
NASA - Leave blank.
NTIS - Leave blank.

Block 13. Abstract. Include a brief *(Maximum 200 words)* factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

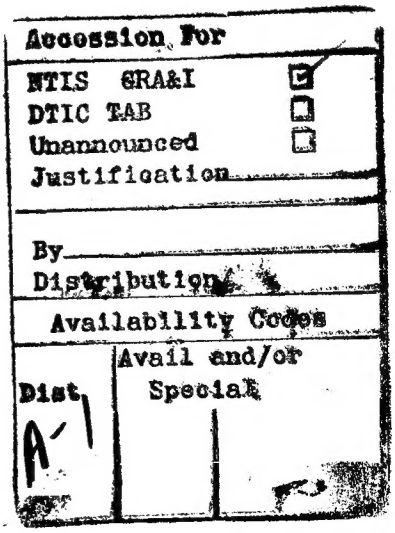
Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code *(NTIS only)*.

Blocks 17. - 19. Security Classifications. Self-explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.

TABLE OF CONTENTS



FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the US Army.

____ Where copyrighted material is quoted, permission has been obtained to use such material.

____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

____ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and Use of Laboratory Animals of the Institute of Laboratory Resources, National Research Council (NIH Publication No. 86-23, Revised 1985).

____ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

PI - Signature 9/29/94 Date

ANNUAL REPORT ON

MIPR NO. 93MM3513

for the period

January 11, 1993 – January 11, 1994

I. INTRODUCTION AND BACKGROUND

The objectives of the above project were formulated in discussion with Mr. Henry Gardner of U.S. Army Medical R&D Command, Ft. Detrick, Maryland. The project purpose and workscope was stated in the proposal as follows: to perform mathematical, statistical and risk-analytical work in support of the mission of the Army Biomedical Research and Development Laboratory (ABRDL). The project continues and extends work performed under MIPR No. 91MM1598.

II. APPROACHES TAKEN AND PROGRESS

Work has been initiated in three areas:

A. *A SIMULATION STUDY OF THE BEHAVIOR OF ESTIMATORS OF THE TERATOGENIC INDEX.*

Appendix A contains a simulation study of the behavior of estimators of the teratogenic index. Approximations for the variance of the teratogenic index and the logarithm of the index are given. Simulation is used to study the behavior of using these approximations to obtain approximate standard errors and confidence intervals for the index. The simulation results suggest that the sampling distribution of the estimator of teratogenic index is not symmetric but the sampling distribution of the logarithm of the estimator is more symmetric. Confidence intervals based on the normal distribution may not have the advertised coverage if the sampling distribution of the statistic is not

symmetric. The simulation results indicate that the coverage of the confidence intervals is reasonable, particularly those based on the logarithm of the index.

B. *DESIGN OF EXPERIMENTS AND ANALYSIS OF THE EXPERIMENTAL DATA.*

Here is an experimental protocol that may well be of considerable usefulness in practice. It is basically the same as that currently used, but recommends an attitude of caution, suspicion, explicitness-of-purpose and search for explanations of what is observed that may verge on the paranoid. The purpose of such an approach is to understand and quantify the sources of variability in the experiment. Dr. Twerdok's establishment of a data base to monitor the health of the medaka will aid in understanding the natural variability of the population.

(a) Choose a number of experimental animals (e.g. medaka fish -- the main consideration -- or rodents) and subject them to specified environmental and dosage conditions. Identify those in "tank" i, $i = 1, 2, \ldots, I$; put $n_i(t_k)$ originally therein: t_k refers to a time of ultimate sacrifice. It will be highly desirable to keep track of happenings in tank i as carefully as possible, e.g. recording temperature measurements, PH, etc., -- even number of fish that die. The initial fish complements of the tanks should be randomized. *Any* extra information about both individual fish or the respective tanks is worth having, as initial variations of same *may* influence the later biological experiences of the fish. Both mean and *variance* of measurements *could* be dose-affected.

(b) Fish *treatment* of interest *may* involve subjecting them to a *steady concentration of a chemical* (DEN, perhaps, or in combination with other affectors) over a *period of time.* At the end of the exposure of the fish they will be removed and examined. Let $n_{ic}(t_k)$ be the number of fish in tank i that have received chemical dosage c constantly over time t_k. Note that the dosage pattern need not be as simple as described; the subscript c simply designates the dosage type administered for the time t_k. The dosage may be time-

varying (bolus, bolus plus constant, constant for time, nothing thereafter, etc.) ... whatever is biologically interesting or meaningful.

(c) *Control* or *reference* treatments are worth having, and often essential if we want to study a more operational (groundwater concentration levels) situation. Unfortunately, these must be carried out in *separate tanks*, since the chemical is in solution. In spite of the exercise of great care there can be between-tanks differences (over and above dosage). Consequently, replication of tank experience is highly advisable, indeed essential, in order to be able to estimate between-tank contribution to variation in endpoints of direct biological concern.

Appendix B reports the analysis of some data from a bioassay study. A procedure to assess the variability between tanks in the same treatment groups is proposed. The procedure is then used to assess the variability between tanks. The analysis suggests that there is a tank effect within treatments.

We will investigate (by mathematics and perhaps computer simulation) the effect of number of replications, numbers of original subjects, dosages, and sacrifice times, plus the various endpoint observations that may be informative, not to mention the influence of the types of parametric dose-response models used to summarize the data *and* the ways those models can be fitted, and the fit quality and informativeness. These studies will contribute to the scientific understanding needed to minimize the number of animals used in experiments.

B.1. Anticipatory Dose-Response Predictor Methodology

Suppose we want to infer the effect of a dosage level of some agent, say TCE, on an endpoint of interest, e.g. occurrence of (pre)cancerous foci. The delay in getting any foci at small doses suggests search for a precursor. One possibility is a *proliferative index* (PI) change or level obtained by a staining technique: roughly speaking the technique identifies the fraction of cells, in a replicative stage at the sampling time. The argument

is that there may be an exploitable relationship between (some form of) PI level, observable relatively soon, and later appearance of foci.

Such an anticipatory strategy is promising and could well be highly profitable for risk analyses. We wish to establish credible analytical tools for handling such data. These tools may well be of use more extensively, to the benefit of risk analysts.

B.2. Models

There are a number of models that may credibly connect PI and F (focii prevalence) data. Here is one such that is essentially off-the-shelf and could be useful.

Logistic

Assume that there is a regression-like relation between

$\underbrace{\text{PI}(t_k)}$	and	$\underbrace{f_\ell / n_\ell}$
Proliferative Index at Sacrifice Time t_k $k = 1, 2, \ldots, \overline{K}$		Fraction of Subjects Exhibiting Response (= Foci) at Sacrifice Time t_ℓ $\ell = 1, 2, \ldots, I$

where t_ℓ is later than $t_{\overline{K}}$.

The basic model is that f_ℓ / n_ℓ may be approximately of the form

$$\frac{\exp\left(a + b_1 \text{PI}(t_1) + b_2 \text{PI}(t_2) + \ldots + b_{\overline{K}} \text{PI}\left(t_{\overline{K}}\right) + c\, t_\ell\right)}{1 + \exp\left(a + b_1 \text{PI}(t_1) + b_2 \text{PI}(t_2) + \ldots + b_{\overline{K}} \text{PI}\left(t_{\overline{K}}\right) + c\, t_\ell\right)}$$

where a and $b_1, b_2, \ldots, b_{\overline{K}}$ and c are unknown constants that can be estimated by maximum likelihood. Here "exp" stands for exponential. This is a standard form for predicting probabilities, available in many statistical packages, such as BMDP and probably SAS, etc.

C. *TOWARDS DECISION-ORIENTED SCORING OF TOXIC-WASTE REPOSITORY/SITE CLEANUP.*

The U.S. continental area is dotted with a number of landsites that have been dedicated to the containment of toxic wastes: so-called toxic waste repositories (TWRs).

Several to many of these sites are located on military bases, e.g., at Aberdeen Proving Ground and at Rocky Mountain Arsenal but elsewhere as well. Such sites contain large amounts of various seriously toxic elements that have been entering ground water and appearing as effluent from the sites. Other environmental components such as soil may well be affected; contaminated soil can be dispersed by wind and rain. Owing in part to the planned reduction of overall U.S. military investment, including requirement for land for weapons testing, but also to a growing appreciation of environmental threat, certain areas containing military (and civilian) TWRs are to be closed to further input and cleaned up so as to reduce hazards to humans in the environment; also of concern is the broader environment as well: its vegetation and wildlife. Note that clean up of such sites that are to be cleaned may not be limited to those that are to be totally closed; the human environmental impact of sites that remain operational will remain of concern.

C.1. The Problem

To clean a site means here to reduce its undesirable impact to a tolerable or inoffensive level, appropriately defined. At least one important component of such impact is measured by a *collection of potentially hazard-related chemical constituents* of the groundwater and effluents associated with the sites. Excessive presence of the above chemicals is believed to be threatening to human life and the environment. Thus clean up is aimed at reducing the concentration of such items in ground water and effluent to a tolerable level.

Clean up is to be accomplished in a cost-effective manner. Note that we use water clean up as an example and would not limit attention to it alone if other elements such as air quality, surface soil condition, *etc.*, are abnormal and are options for improvement or "clean up," as will very likely be the case.

There are various problems of detail that confront a decision-maker who deals with clean up. We review these as they are now understood.

Some specific problems are these.

(1) The contents, and pollution potential, of each TWR are only vaguely known. The potential for environmental damage may be related to complex combinational or serendipidous behavior of many chemicals.

(2) The concentration of the various potentially hazardous components in groundwater and effluent are likely to vary haphazardly ("randomly") over time due to local conditions such as rainfall and general groundwater level, but also because of the rate of decomposition of materials and their containments in the ground, chemical reactions, *etc.* This noisy background helps to obscure desired signals of a concentration decrease resulting from clean up effort. In fact, it has been remarked by Travis (1992) that pumping to remediate groundwater in contaminated aquifers may be only apparently, and actually just temporarily, effective. The reason is that while pumping may reduce concentrations, dense NAPLs are nearly permanently in place at aquifer bottoms, where their dissipation is by slow molecular diffusion. Observed contaminant concentration reduction while pumping is largely the effect of dilution; concentrations have been observed to go back up once pumping stops. If soil is contaminated for a long time with hydrophobic chemicals the same effect prevails.

(3) The relative importance of clean up of the various items (chemicals) may be unclear; some may require more attention than others. In fact "clean up" needs to be *defined* in a way that is agreed upon, scientifically supportable (or not wholly specious), cost-effective and practical, and communicable.

(4) Chemical indicators alone may well not portray hazard adequately, particularly as these conspire to affect complex biological organisms such as mankind, wildlife, vegetation, *etc.* For this reason, testing for clean up with actual organisms, e.g., the Japanese medaka fish or frog embryos, is an attractive supplement. Fish may be a good medium by which to track a propensity for certain diseases such as cancer, but may well be useless for other indications, such as air quality. Other biological markers, such as

plants, have promise. The use of organism indicators sounds primitive, but has been historically effective. These biomonitoring test systems (BTS) perform as interpretors of complex dosage. However it should be noted and considered that *not all individual biological markers are identical*. Differences between individual organisms, e.g., fish, may well mask responses to different contaminant levels, producing a "noisy signal" concerning current, or average, contaminant level. It is imperative that careful and appropriate analytical statistical tools be brought to bear to guide acquisition and interpretation of data from TWR monitoring. A further issue is how to combine information from analyses of individual BTS (applied at different locations and times at a specified site).

C.2. Decision Assistance

A decision maker has various options with respect to a TWR. Here are some.

(a) Leave the TWR alone. This may be acceptable in certain cases in view of cost and judged impact; see Travis (1992). Or it may reflect a priority scheme that postpones action in favor of a more pressing need elsewhere. An assessment procedure that reliably and defensibly assesses the future effect seems necessary. Expert judgement is useful but not sufficient.

(b) Isolate or contain without clean up. This may have been done in Europe (Germany) with certain rivers, e.g., the Rhine. A monitoring and assessment procedure seems essential.

(c) Complete surgical clean up once and for all by excavation, offsite re-location, or decomposition of contents. Replace soil. This "organ transplant" (Oregon transplant!) procedure may be far too expensive in practice, but is perhaps an ideal. Once again, an attempt to measure and quantitatively assess the degree of cleanup is desirable.

(d) Perform *partial clean up* as in (c), then *process* water that interacts with TRW and reaches outside, e.g., enters groundwater that is used elsewhere, or flows into streams or rivers. This may be a common option. Its effectiveness may well depend on contaminants present.

Question: How does the decision maker decide whether the processing procedure is sufficient; i.e., is the permitted output *clean enough*? The answer to this question must, for political reasons, be defensible on a level somewhat comprehensible to an informed

attentive layman. Desirably, the procedure adopted must also be legally defensible and cost effective. A lucid and objective quantitative approach seems essential. We are continuing to actively review and appraise the relevant literature and directives, e.g., from EPA.

In what follows we propose quantitative attacks, particularly on option (d), partial clean up and processing of effluent.

C.3. Quantitative Approaches; States

The decision maker is potentially able to measure the current (time t) concentrations of n individual chemical contaminants at chosen sampling times t; call these

$$y_1(t), y_2(t), ..., y_n(t). \tag{1}$$

For instance the first component $y_1(t)$ might be ppm of arsenic as measured at $t = 15$ days after clean up begins; the last component, $y_n(t)$, might be a concentration of dioxin. It is expected that, before treatment, at $t = 0$, $y_i(t)$, $t < 0$, will vary haphazardly, possibly because of variations attributable to season of the year, basic water flow, temperature, age of certain TWR contents, and so on; in other words each concentration $y_i(t)$, $t > 0$, is a **time series**, and the collection of all is a multi-variable time series. Ideally we intend and anticipate that the general level (e.g., mean) of such time series will decrease with time measured from the "instant" when processing starts; presumably we also want large excursions, or pulses, of contaminant concentration to become much less frequent as treatment continues. Note that the above concentrations may realistically be composites of measures taken at various spatial points in an aquifer, so a more inclusive portrayal of reality is the time-space series $(y_{ij}(t), \ell_i)$, where ℓ_i is the location of the observation j. The subsequent discussion omits consideration of this detail.

In addition to chemicals we include organic indicators (biomarkers) such as the numbers of medaka livers containing tumors out of a number exposed in a sample and/or the result of FETAX assays; call the results of these assays

$$z_1(t), z_2(t), ..., z_m(t) \quad (2)$$

if there are m such indicators. An important question is the choice of the number of biomarkers to use and the frequency of their use. Note that the chemical concentrations $\underline{y}(t)$ will probably be modeled on "continuous" random processes (possibly, but not necessarily, Gaussian), while the counts will be "discrete" (i.e. taking on values like 0, 1, 2, ..., 13, ...). Of course all organic measures (we allow for $m \geq 1$) need not be discrete. Thus we think of the status ("state") of the system (TWR) to be given by $(\underline{y}(t), \underline{z}(t))$ at time t, i.e. two collections of measurements or counts. An important and practical research question is to specify a suitable, and cost-effective, *contamination profile or site state vector* $(y(t), z(t))$.

C.4. Quantitative Assessment of Clean-up Adequacy

An (impractical) ideal would be to insist on, and attempt to guarantee, *no* (zero) contamination by each identified contaminant after the cleanup project begins. This is unrealistic because, first, true concentrations are likely to fluctuate over time and over space, i.e. with location within and near the site, and second, *measured* concentrations, or their surrogates such as the number of tumor-infected medaka livers in an exposed group of fish will not give a totally noise-free indication of the true effective concentration prevailing at a particular time. In other words, a measurement portrait may very likely be a somewhat inaccurate portrayal of a particular item's true time-space concentration. With this in mind consider the following

Objective: Set up a *simple index* of the TWR's overall contamination condition at or close to, any time t that accounts for the effects of measurement error, the inherent variability of biological systems, and estimated true-value fluctuation for each (recognized) contamination component.

It cannot now be argued that a simple single index can be devised that summarizes "site health" in a totally satisfactory and non-controversial way. However, effort should go towards devising such an index, and establishing its credibility if only to

assist in communication and to guide policy and decision makers. Behind such an index should be more specific measures of cleanup performance, i.e. response to cleanup efforts targeting specific pollutants.

Some tentative suggestions for accomplishing the objective are made below. There are various options that have different good and bad points, not all of which are well-understood. It is proposed to lay out some of these options and to continue to conduct research on their relative merits. It is also proposed to search for and evaluate other options.

Option 1: Multiple Hypothesis Testing

A conventional way to assess the effect of a treatment is to choose a relevant measurable response whose generic value is $y_i(t)$ for the i^{th} contaminating element at time t, measure it (replicate) J times under (a) remediated and (b) control conditions, or alternatively, (b'), with reference to a tolerable threshold y_i^*. Then perform a classical one-sided hypothesis test of a suitable null hypothesis. Suppose a relevant test statistic is denoted by

$$\Delta_i(t) = \overline{y}_i(t;(a)) - \overline{y}_i(t;(b)) \tag{3}$$

where $\overline{y}$ denotes a summary of the J replicates mentioned earlier; this summary can be a simple mean, a robust alternative (e.g., median or other M-estimate), or a relevant parameter estimated by likelihood or Bayes methodology. Assume that $\overline{y}$ responds positively to the presence of contaminant: the greater the concentration of element i present, the greater would Δ_i tend to be. Then an appropriate null hypothesis *could* be that $\Delta_i(t)$ is a sample from a Normal distribution with zero mean and standard deviation σ_i. A test of the i^{th} null hypothesis alone would be: reject it, i.e. the hypothesis that remediation has brought the concentration of element i under control, if $\Delta_i(t) \geq \tilde{\Delta}_i$, where the cutoff value $\tilde{\Delta}_i$, is, say, a 95^{th} percentile of a normal distribution with mean zero and standard deviation σ_i, or better, the corresponding t-distribution.

Unfortunately if the number of contaminant elements being tested for is large, it becomes likely that *at least one* such test asserts a significant difference, namely that $\Delta_i(t) \geq \tilde{\Delta}_i$, even if all null hypotheses are actually true. This is the **multiplicity problem.** A way of addressing it is as follows. A *p*-value associated with $\Delta_i(t)$ is the probability that any random sample from the above population exceeds the observed/measured $\Delta_i(t)$-value; let p_i be the numerical value of the i^{th} *p*-value. Actually, since σ_i will be unknown and hence must be estimated the appropriate reference distribution should be a Student *t*. Note that this *p*-value number becomes small if the difference between the responses under remediated conditions and under control or threshold conditions is large, indicating that remediation is not (yet) effective. If there are *I* (e.g., 10) different contaminating elements being tested for, then one can assess for the *combined significance* of all elements by use of the Fisherian statistic

$$\chi^2 = -\sum_{i=1}^{I} \ln(p_i). \tag{4}$$

Under the null hypothesis of no difference, in any element, between remediated response and control the above is distributed as a *chi*-square random variable with $2I$ degrees of freedom; an observed value high enough to exceed the 95^{th} or 99^{th} percentile of such a *chi*-square distribution suggests that, overall, the current remediation effort has not been successful. A useful informal supplement would be to plot the individual p_i-values to see if they appear to be uniformly distributed over [0,1]; if most were close to unity, but several close to zero, the latter "several" would be implicated as those elements not yet affected by remediation.

Note that the above process is suggested as an informal screening procedure. It has various flaws; many are identified in the *NRC Combination of Information Report* (1992), abbreviated NRC/CI. For one thing the test has no explicit dependence on sample size for the individual element summaries; for another, there is no acknowledged dependence on the individual test statistic distributions, nor on the cost of either

erroneously accepting the remediation as complete when it is not or of continuing with the effort when it is not required. For still another, there is no attempt to "borrow strength" by utilizing information about the effect of the same remediation strategy on the same contaminating elements at other toxic waste repositories.

Option 2: Maximum Test Risk

For the k^{th} biomonitoring or other test system ($k = 1, \ldots, K$) determine the largest dose level d_k (e.g. lowest groundwater dilution) so that the probability that the response at that dose is greater than that for the control is less than a (small) number r. A possible decision rule is to declare the smallest value (lowest concentration) $\min_k d_k$ safe at maximum test risk level r.

III. RECOMMENDATION

Further work, both theoretical and applied, is required to put the above ideas for combining information into practice. It is proposed that this work and research into the quantitative analysis of bioassay data be continued.

BIBLIOGRAPHY

Schwartz, F. W. (Chairman) (1990). *Ground Water Models: Scientific and Regulatory Applications*. National Research Council. National Academy Press, Washington, DC.

van der Heijde, P. K. M., A.I. El-Kadi and S. A. Williams (1988). *Groundwater Modeling: An Overview and Status Report*. Report Number GWMI 88-10. International Groundwater Modeling Center. Colorado School of Mines; Institute for Ground-Water Research and Education. Golden, Colorado, 80401-1887.

Anderson, M. P. and W. W. Woessner (1991). *Applied Groundwater Modeling*. Academic Press.

Moolgavkar, S. H. and D. J. Venzon (1979). "Two-event models for carcinogenesis; incidence curves for childhood and adult tumors." *Math. Biosciences*, 47, pp. 55-79.

Gaver, D. P. (Chairman), D. Draper, P. K. Goel, J. B. Greenhouse, L. V. Hodges, C. N. Morris, and C. Waternaux (1992). *On Combining Information. Statistical Issues and Opportunities for Research*. Report of Panel on Statistical Issues and Opportunities for Research in the Combination of Information. Committee on Applied and Theoretical Statistics; Board on Mathematical Sciences, National Research Council, National Academy Press.

Gaver, D. P. (1992). Modeling and statistical analysis of bioassay data. Final report for MIPR No. 91MM1598. Dept. of Operations Research, Naval Postgraduate School, Monterey, CA 93943. (with Prof. P. A. Jacobs).

Travis, C. C. (1992). "Toxic waste in groundwater: can it be removed?" *The J. of NIH Research*, vol. 4; pp. 49-51.

Morgan, B. J. T. (1992). *Analysis of Quantal Response Data*. Chapman & Hall, London.

Collett, D. (1991). *Modeling Binary Data*. Chapman & Hall, London.

Efron, B., and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, London.

Twerdok, L. E. (1994). "Monitoring the health status of medaka used in toxicology studies." Talk presented at Sixth Annual Research Review, September 20–21, 1994, Frederick, MD.

Appendix A

A Simulation Study of the Behavior of Estimators of the Teratogenic Index

Animal experiments are used to study the effects of the dose of a potential toxin. One measure of the effect is the dose which is lethal to 50% of the population LD50 = μ_N. Another measure is the dose which produces undesirable symptoms in 50% of the population ED50 = μ_D. A combined measure of the effect of the toxin is the ratio $\gamma = \mu_N/\mu_D$ of the two doses; such a ratio is the *teratogenic index*. Large values of $\gamma >> 1$ are of concern.

In this note we use simulation to investigate the behavior of two approximate expressions for the standard deviation of $\hat{\gamma} = \hat{\mu}_N / \hat{\mu}_D$.

Two Approximate Expressions for the Variance of $\hat{\gamma}$ and log $\hat{\gamma}$.

In this section we use the "delta method" to obtain an approximation for the variance of the ratio $\hat{\gamma} = \hat{\mu}_N / \hat{\mu}_D$ and an approximation for the variance of the log ratio $\log\hat{\gamma}$. This is a simple way of combining standard errors from dose-response (e.g. probit) analyses to obtain approximate standard errors and confidence intervals for γ.

The probit model is often used to estimate μ_N and μ_D. In this case, the sampling distribution of $\hat{\mu}_N$ and $\hat{\mu}_D$ is asymptotically normal. Hence we write

$$\boxed{\hat{\gamma} = \hat{\mu}_N / \hat{\mu}_D} \qquad (1)$$

$$= \frac{\mu_N + \varepsilon_N}{\mu_D + \varepsilon_D}$$

$$\approx \mu_N(1 + \varepsilon_N / \mu_N)\left[\frac{1}{\mu_D} - \frac{1}{\mu_D^2}\varepsilon_D\right]$$

where μ_N and μ_D are the true parameter values and ε_N and ε_D are independent normal random variables with mean 0; the last expression follows from a partial Taylor expansion of $(\mu_D + \varepsilon_D)^{-1}$ about μ_D. Thus

$$\hat{\gamma} \approx \frac{\mu_N}{\mu_D}(1+\varepsilon_N/\mu_N)(1-\varepsilon_D/\mu_D)$$
$$\approx \gamma\left[1+\frac{\varepsilon_N}{\mu_N}-\frac{\varepsilon_D}{\mu_D}\right]$$

Hence,

$$Var\ \hat{\gamma} \approx \gamma^2\left[\frac{\sigma_N^2}{\mu_N^2}+\frac{\sigma_D^2}{\mu_D^2}\right]$$

where σ_N^2 (respectively σ_D^2) is the variance of ε_N (respectively ε_D). A crude but convenient estimate of the Var $\hat{\gamma}$ is

$$\hat{Var}\ \hat{\gamma} \approx \hat{\gamma}^2\left[\frac{s_N^2}{\hat{\mu}_N^2}+\frac{s_D^2}{\hat{\mu}_D^2}\right] \tag{2}$$

where s_N^2 and s_D^2 are the sample variances of μ_N and μ_D. The *standard error* of $\hat{\gamma}$ is

$$\boxed{SE[\hat{\gamma}] = \hat{\gamma}\sqrt{\frac{s_N^2}{(\hat{\mu}_N)^2}+\frac{s_D^2}{(\hat{\mu}_D)^2}}.} \tag{3}$$

The *logarithm* of a ratio estimate, such as γ, is often more stable numerically and often has a more symmetric sampling distribution than the sampling distribution of the ratio itself. We next derive an approximate expression for the $Var[\log\hat{\gamma}]$. Note that

$$\boxed{\log\hat{\gamma} = \log\hat{\mu}_N - \log\hat{\mu}_D} \tag{4}$$
$$= \log(\mu_N+\varepsilon_N) - \log(\mu_D+\varepsilon_D)$$
$$\approx \log\mu_N + \frac{1}{\mu_N}\varepsilon_N - \log\mu_D - \frac{1}{\mu_D}\varepsilon_D$$

where the last expression follows from the first two terms of a Taylor expansion of the log about μ_N and μ_D. Thus,

$$Var\ [\log \hat{\gamma}] \approx \left[\frac{\sigma_N^2}{\mu_N^2} + \frac{\sigma_D^2}{\mu_D^2}\right] \tag{5}$$

A crude estimate of $Var[\log \hat{\gamma}]$ is

$$\hat{Var}\ [\log \hat{\gamma}] \approx \left[\frac{s_N^2}{\hat{\mu}_N^2} + \frac{s_D^2}{\hat{\mu}_D^2}\right] \tag{6}$$

$$\boxed{SE[\log \hat{\gamma}] = \hat{\gamma}\sqrt{\frac{s_N^2}{\hat{\mu}_N^2} + \frac{s_D^2}{\hat{\mu}_D^2}}\ .} \tag{7}$$

A Simulation Experiment.

In the following simulation experiments $s_N = \frac{0.4}{\sqrt{10}}$ and $s_D = \frac{0.3}{\sqrt{10}}$ are fixed. These particular numbers were obtained from a manuscript by F. Hoffman. For the k^{th} replication of the experiment, a random number $\hat{\mu}_N(k)$ (respectively $\hat{\mu}_D(k)$) is drawn from a normal distribution with mean μ_N and variance s_N^2 (respectively μ_D and s_D^2). The estimates $\hat{\gamma}(k) = \hat{\mu}_N(k) / \hat{\mu}_D(k)$ and $\log \hat{\gamma}(k) = \log \hat{\mu}_N(k) - \log \hat{\mu}_D(k)$ are computed. The sample asymptotic standard deviations which are the square roots of (1) and (2)

$$\hat{v}_\gamma(k) = \left(\frac{\hat{\mu}_N(k)}{\hat{\mu}_D(k)}\right)\left(\frac{s_N^2}{\hat{\mu}_N(k)^2} + \frac{s_D^2}{\hat{\mu}_D(k)^2}\right)^{1/2}$$

and

$$\hat{v}_{\log \gamma}(k) = \left[\frac{s_N^2}{\hat{\mu}_N(k)^2} + \frac{s_D^2}{\hat{\mu}_D(k)^2}\right]^{1/2}$$

are evaluated. Approximate 95% normal confidence intervals are calculated

$$I_\gamma(k) = \frac{\hat{\mu}_N(k)}{\hat{\mu}_D(k)} \pm (1.96)\,\hat{v}_\gamma(k)$$

$$I_{\log \gamma}(k) = \log\left[\frac{\hat{\mu}_N(k)}{\hat{\mu}_D(k)}\right] \pm (1.96)\,\hat{v}_{\log \gamma}(k)$$

The simulation is replicated 500 times and the following statistics are computed. The sample means of the estimated ratio and the estimated log ratio

$$\bar{\gamma} = \frac{1}{500}\sum_{k=1}^{500} \hat{\mu}_N(k)/\hat{\mu}_D(k) \equiv \frac{1}{500}\sum_{k=1}^{500} \hat{\gamma}(k)$$

$$\overline{\log \gamma} = \frac{1}{500}\sum_{k=1}^{500} \log \hat{\mu}_N(k) - \log \hat{\mu}_D(k) \equiv \frac{1}{500}\sum_{k=1}^{500} \log \hat{\gamma}(k)$$

and the sample standard deviations of the estimated ratio and estimated log ratio

$$\hat{\sigma}_\gamma = \left[\frac{1}{499}\sum_k \left(\hat{\gamma}(k) - \bar{\gamma}\right)^2\right]^{1/2}$$

$$\hat{\sigma}_{\log \gamma} = \left[\frac{1}{499}\sum_k \left(\log \hat{\gamma}(k) - \overline{\log \gamma}\right)^2\right]^{1/2}.$$

The average length of the confidence intervals of the ratio and log ratio are computed where

$$\bar{L}_\gamma = \frac{1}{500}\sum_k L_\gamma(k)$$

with

$$L_\gamma(k) = 2(1.96)\hat{v}_\gamma(k)$$

and

$$\bar{L}_{\log \gamma} = \frac{1}{500}\sum_k L_{\log \gamma}(k)$$

with

$$L_{\log \gamma}(k) = 2\exp\left\{(1.96)\hat{v}_{\log \gamma}(k)\right\};$$

the two endpoints of the confidence interval for $\log \gamma$ are exponentiated to give an interval for γ. Finally the fraction of intervals $I_\gamma(k)$ which cover μ_N/μ_D and the fraction of intervals $I_{\log\gamma}(k)$ which cover $\log \mu_N/\mu_D$ are computed.

The results are presented in Table 1. Displayed in Table 1 are the asymptotic standard deviation in each case.

$$v_\gamma = \frac{\mu_N}{\mu_D}\left[\frac{s_N^2}{\mu_N^2} + \frac{s_D^2}{\mu_D^2}\right]^{1/2}$$

$$v_{\log\gamma} = \left[\frac{s_N^2}{\mu_N^2} + \frac{s_D^2}{\mu_D^2}\right]^{1/2}$$

and the corresponding sample standard deviations $\hat{\sigma}_\gamma$ and $\hat{\sigma}_{\log\gamma}$. Comparison of v_γ and $\hat{\sigma}_\gamma$ suggests that the approximate standard deviation for $\hat{\gamma}$ is reasonably accurate as long as μ_D is not too close to 0. Similarly comparison of $v_{\log\gamma}$ and $\hat{\sigma}_{\log\gamma}$ suggests that the approximate variance for $\log\hat{\gamma}$ is even more accurate. Histograms of the estimates $\{\hat{\gamma}(k)\}$ and $\{\log\hat{\gamma}(k)\}$ are displayed as Figure 1 for the case $\mu_N = 1$, $\mu_D = 0.5$ in which μ_D is close to 0. Note that the histogram of $\{\hat{\gamma}(k)\}$ suggests that the sampling distribution of γ is somewhat skewed to the right. The histogram of $\{\log\hat{\gamma}(k)\}$ appears more symmetric. Confidence intervals based on the normal distribution may not have the advertised coverage if the sampling distribution of the statistic is not symmetric.

Also displayed in Table 1 are the fraction of simulation confidence intervals that cover the true $\gamma = \mu_N/\mu_D$ and the sample mean of the lengths of the simulated confidence intervals. Recall that the endpoints of the simulated confidence intervals for $\log\gamma$ are exponentiated before computing the length and computing the sample mean. As a result the sample means of the length of the confidence interval for γ and $\log\gamma$ are comparable. The results indicate that the coverage of the confidence intervals is reasonable. However, the average length of the interval can be large. It is particularly large (~2) when $\mu_N = 1$, $\mu_D = 0.5$ and $\gamma = 2 = 1/0.5$, i.e. when the denominator is small.

Simulation Study of Estimates of the Standard Deviation of the Teratogenic Index

$$s_N = 0.4/\sqrt{10} \qquad s_D = 0.3/\sqrt{10} \qquad \hat{\gamma} = \hat{\mu}_N / \hat{\mu}_D$$

μ_N	μ_D	γ μ_N/μ_D	Nbr of Repl.	Asym std dev for γ v_γ	Sample std dev for γ $\hat{\sigma}_\gamma$	Asym std dev for log γ $v_{\log\gamma}$	Sample std dev for log γ $\hat{\sigma}_{\log\gamma}$	Fract 95% CI covering (average width)* γ	log γ
1	0.5	2	500	0.46	0.51	0.23	0.23	0.94 (1.99)	0.96 (2.07)
2	1	2	500	0.23	0.25	0.11	0.12	0.95 (0.92)	0.95 (0.93)
1.5	1	1.5	500	0.19	0.20	0.13	0.13	0.94 (0.76)	0.94 (0.77)
3	1	3	500	0.31	0.31	0.10	0.10	0.96 (1.25)	0.96 (1.26)

* average width is for a CI for γ. The endpoints of the CI for log γ are exponentiated.

SIMULATION ESTIMATES OF TERATOGENIC INDEX;MUN=1 MUD=0.5

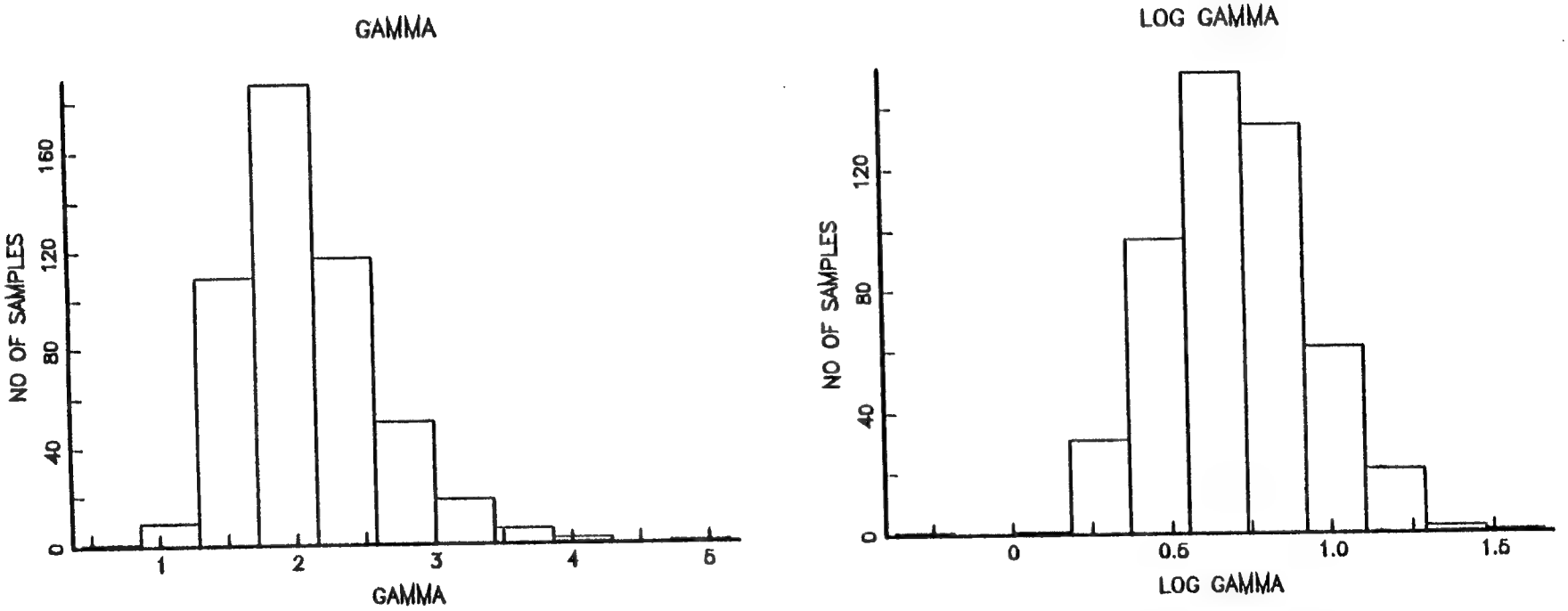


FIGURE 1

Appendix B

MEGA MEDAKA STUDY
A Study of Tank Variability

Medaka fish are exposed to different levels of a (potential) toxin. Each treatment including a control has $N_T = 4$ tanks allocated to it. Some of the fish in each tank are sacrificed at 4, 6, and 9 months after their initial exposure. Their livers are examined and the number of fish that have hepatocellular neoplasms and/or carcinomas are recorded.

The purpose of this note is to study the variability of the tanks within a treatment level.

The simplest model is that the fish in all tanks within a treatment are subject to the same environment. If $X_{i,e}(s)$ is the number of fish in tank i in treatment e at time s whose livers have neoplasms or carcinomas out of the $N_{i,e}(s)$ that were sacrificed, then the simplest model is $\{X_{i,e}(s), i = 1, \ldots, N_T\}$ are independent binomial random variables; $X_{i,e}(s)$ has a binomial distribution with $N_{i,e}(s)$ trials and probability of occurrence of neoplasm or carcinoma $p_e(s)$. For this model the maximum likelihood estimate of $p_e(s)$

$$\hat{p}_e(s) = \frac{\sum_{i=1}^{N_T} x_{i,e}(s)}{\sum_{i=1}^{N_T} n_{i,e}(s)}$$

which has asymptotic variance

$$\hat{V}_e(s) = \frac{\hat{p}_e(s)[1-\hat{p}_e(s)]}{\sum_{i=1}^{N_T} n_{i,e}(s)}.$$

This model has implications about how much variability the results from each tank can display.

To study the tank variability the following calculations are performed within each treatment group at each time s. Each tank was left out in turn and $\hat{p}_e(s)$ was estimated using the $n_I(s)$ fish sacrificed from the remaining 3 tanks. The following simulation is then conducted. The i^{th} replication consists of the following. A binomial random number (O) with $n_I(s)$ trials and probability of success $\hat{p}_e(s)$ is drawn and the fraction $p_O = O/n_I(s)$ is computed. If the left out tank has $n_O(s)$ fish sacrificed from it, then the 0.05 and 0.95 percentiles, $q_L(i)$ and $q_U(i)$, from a binomial distribution with $n_O(s)$ trials and probability of success p_O are found. The simulation is replicated N_r times. A confidence interval for the fraction of sacrificed fish in the left out tank which exhibit neoplasms/ carcinomas is

$$\left[\frac{\sum_{i=1}^{N_r} q_L(i)}{N_r}, \frac{\sum_{i=1}^{N_r} q_U(i)}{N_r}\right]$$

The observed fraction is then compared to this interval. If the model is correct, then the observed fraction should fall into the confidence interval the majority of the time. Results appear in Table 1 for that part of the experiment which uses fish that are 6 days of age at the start of the experiment and Table 2 for that part of the experiment which uses fish that are 52 days of age at the start of the experiment.

The results suggest that there is more variability between tanks for that part of the experiment that involves fish that are 6 days of age at the start of the test. The results suggest that when fitting parametric models to the data, a variable for tank effect be included in the model.

Table 1

Fish at 6 days of Age at Start of Test
(500 replications)

Sacrifice Time		4 months		6 months		9 months	
Treat-ment (mg/ℓ)	Tank Left Out	Conf Interval	Fraction Abnormal in Left Out Tank	Conf Interval	Fraction Abnormal in Left Out Tank	Conf Interval	Fraction Abnormal in Left Out Tank
Control	1	[0,0]	0*	[0,0]	0*	[0.008,0.09]	0
	2	[0,0]	0*	[0,0]	0*	[0,0]	0.18
	3	[0,0]	0*	[0,0]	0*	[0.007,0.08]	0.08*
	4	[0,0]	0*	[0,0]	0*	[0.01,0.11]	0
2.5	1	[0.001,0.03]	0	[0.04,0.15]	0.08*	[0.20,0.40]	0.19
	2	[0.001,0.03]	0	[0.01,0.08]	0.24	[0.20,0.40]	0.09
	3	[0,0]	0.04	[0.05,0.17]	0.04	[0.12,0.31]	0.42
	4	[0.001,0.04]	0	[0.06,0.19]	0	[0.15,0.36]	0.29*
5.0	1	[0.01,0.08]	0	[0.13,0.28]	0.04	[0.45,0.69]	0.32
	2	[0.001,0.03]	0.08	[0.10,0.25]	0.12	[0.34,0.58]	0.61
	3	[0.005,0.06]	0.04*	[0.08,0.20]	0.24	[0.41,0.64]	0.43*
	4	[0.01,0.08]	0	[0.07,0.20]	0.25	[0.33,0.57]	0.67
10.0	1	[0.04,0.15]	0.12	[0.32,0.51]	0.20	[0.63,0.85]	0.75*
	2	[0.03,0.13]	0.16	[0.24,0.42]	0.44	[0.62,0.83]	0.81*
	3	[0.07,0.20]	0	[0.29,0.48]	0.28	[0.68,0.88]	0.64
	4	[0.04,0.15]	0.12*	[0.22,0.39]	0.52	[0.63,0.84]	0.79*

* Fraction of fish with neoplasms/carcinomas for left out tank falls within the confidence interval computed with the other tanks.

Table 2

Fish at 52 days of Age at Start of Test
(500 replications)

Sacrifice Time		4 months		6 months		9 months	
Treat-ment (mg/ℓ)	Tank Left Out	Conf Interval	Fraction Abnormal in Left Out Tank	Conf Interval	Fraction Abnormal in Left Out Tank	Conf Interval	Fraction Abnormal in Left Out Tank
Control	1	[0.006,0.06]	0	[0,0]	0*	[0.001,0.04]	0.04*
	2	[0.001,0.04]	0.04*	[0,0]	0*	[0.005,0.06]	0
	3	[0.005,0.06]	0	[0,0]	0*	[0.005,0.06]	0
	4	[0.001,0.03]	0.04	[0,0]	0*	[0.001,0.04]	0.04*
2.5	1	[0,0]	0.04	[0.005,0.06]	0.08	[0.03,0.13]	0.32
	2	[0.001,0.03]	0	[0.02,0.10]	0	[0.10,0.25]	0.04
	3	[0.001,0.03]	0	[0.02,0.10]	0	[0.07,0.21]	0.13*
	4	[0.001,0.03]	0	[0.005,0.06]	0.08	[0.09,0.24]	0.05
5.0	1	[0,0]	0*	[0.05,0.17]	0.16*	[0.05,0.18]	0.17*
	2	[0,0]	0*	[0.06,0.18]	0.12*	[0.06,0.18]	0.18*
	3	[0,0]	0*	[0.05,0.17]	0.16*	[0.06,0.19]	0.13*
	4	[0,0]	0*	[0.08,0.21]	0.04	[0.09,0.23]	0.04
10.0	1	[0.01,0.08]	0	[0.12,0.26]	0.12*	[0.20,0.38]	0.33*
	2	[0.005,0.06]	0.04*	[0.11,0.25]	0.16*	[0.26,0.45]	0.14
	3	[0.006,0.06]	0.04*	[0.10,0.24]	0.16*	[0.16,0.34]	0.43
	4	[0.005,0.06]	0.04*	[0.08,0.21]	0.24	[0.22,0.41]	0.27*
20.0	1	[0.05,0.17]	0	[0.24,0.42]	0.24*	[0.51,0.71]	0.59*
	2	[0.03,0.13]	0.08*	[0.20,0.37]	0.40	[0.48,0.70]	0.65
	3	[0.02,0.10]	0.17	[0.21,0.38]	0.36*	[0.53,0.73]	0.55*
	4	[0.03,0.13]	0.08*	[0.25,0.43]	0.24	[0.50,0.70]	0.63

* Fraction of fish with neoplasms/carcinomas for left out tank falls within the confidence interval computed with the other tanks.